Experimenters' Influence on Mental-Imagery based Brain-Computer Interface User Training

Léa Pillette*
Inria, LaBRI (Univ. Bordeaux, CNRS, Bordeaux-INP),
200 av. de la Vieille Tour, 33400 Talence, France
lea.pillette@ensc.fr

Aline Roc*
Inria, LaBRI (Univ. Bordeaux, CNRS, Bordeaux-INP),
200 av. de la Vieille Tour, 33400 Talence, France
aline.roc@inria.fr

Bernard N'Kaoua
Handicap, Activity, Cognition, Health, Inserm / University of Bordeaux,
146 rue Léo Saignat, Bât 1B, étage 2, 33076 Bordeaux cedex, France
bernard.nkaoua@u-bordeaux.fr

Fabien Lotte
Inria, LaBRI (Univ. Bordeaux, CNRS, Bordeaux-INP),
200 av. de la Vieille Tour, 33400 Talence, France
fabien.lotte@inria.fr

*: co-first authorship. Both authors contributed equally.

Corresponding author: Léa Pillette (lea.pillette@ensc.fr, +33 6 77 96 81 92)

Quantitative data of the manuscript

Word count for the abstract: 222 words. **Word count for the text:** 8509 words. **Character count for the title:** 88 characters including spaces, punctuation and subtitle. **Number of references:** 54. **Number of tables:** 1. **Number of figures:** 6. **Number of appendix:** 3.

Funding information

This work was supported by the French National Research Agency (project REBEL, grant ANR-15-CE23-0013-01) and the European Research Council with the Brain-Conquest project (grant ERC-2016-STG-714567).

Declarations of interest

None.

Abstract

Context

Motor Imagery based Brain-Computer Interfaces (MI-BCIs) enable their users to interact with digital technologies, e.g., neuroprosthesis, by performing motor imagery tasks only, e.g., imagining hand movements, while their brain activity is recorded. To control MI-BCIs, users must train to control their brain activity. During such training, experimenters have a fundamental role, e.g., they motivate participants. However, their influence had never been formally assessed for MI-BCI user training. In other fields, e.g., social psychology, experimenters' gender was found to influence experimental outcomes, e.g., behavioural or neurophysiological measures.

Objective

Our aim was to evaluate if the experimenters' gender influenced MI-BCI user training outcomes, i.e., performances and user-experience.

Methods

We performed an experiment involving 6 experimenters (3 women) each training 5 women and 5 men (60 participants) to perform right versus left hand MI-BCI tasks over one session. We then studied the training outcomes, i.e., MI-BCI performances and user-experience, according to the experimenters' and subjects' gender.

Results

A significant interaction between experimenters' and participants' gender was found on the evolution of trial-wise performances. Another interaction was found between participants' tension and experimenters' gender on the average performances.

Conclusion

Experimenters' gender could influence MI-BCI performances depending on participants' gender and tension.

Significance

Experimenters' influence on MI-BCI user training outcomes should be better controlled, assessed and reported to further benefit from it while preventing any bias.

Keywords

1. Brain-Computer Interfaces
2. Mental imagery
3. User training
4. Experimenter influence
5. Gender

1 Introduction

Motor Imagery based Brain-Computer Interfaces (MI-BCIs) enable their users to send commands to external digital devices by performing motor imagery tasks only, e.g., imagining hands or feet movements, while their brain activity is being recorded [9]. The system has to estimate the motor imagery task that the users perform from the variations occurring in their brain activity, often recorded using electroencephalography (EEG).
The MI-BCI technology has promising medical applications. For instance, BCIs based on motor imagery and motor attempt were used for motor rehabilitation after stroke [1, 52]. They can also be used to control interfaces of communication [2], which is particularly useful for patients with limited or complete loss of the functional ability to communicate caused by a severe loss of voluntary muscular control [2]. MI-BCIs are also used for non-medical applications. For instance, they represent a new tool to control video-games [20].

1.1 Brain-Computer Interfaces user training

Before being operational, MI-BCIs require that both the computer and the user learn during dedicated training phases [9]. On the one hand, the computer must learn to recognize the variations occurring in users' brain activity while they perform the different mental-imagery tasks. On the other hand, the users must learn to produce a stable and distinguishable pattern of brain activity for each of the commands that they wish to send to the computer [27]. Both the computer training and the user training are highly interdependent but the user and the computer are generally trained separately, which probably partly explain the lack of reliability of the system [35].
This current lack of reliability of the system limits the development of MI-BCI applications. Indeed, 10 to 30% of naive users cannot control MI-BCIs, even after some training [30]. There are several lines of research aiming at improving the efficiency of MI-BCIs. Most focus on the improvement of machine learning methods, see, e.g., [17, 23]. A few also focus on the improvement of the user training. Indeed, it has been shown both theoretically and experimentally that current user training approaches may not allow all users to acquire the skills necessary to use MI-BCIs [15, 25].
During their training, users train to control a feedback representing what the computer recog-

nizes of the mental task that they are performing. A feedback is an information which is provided to a learner regarding aspects of the performance or understanding of the task/skills to learn [12]. It is a fundamental component of the MI-BCI training [25]. Several research were led in order to improve the feedback [37], for instance by using more realistic cues [34, 47]. Users could need specific feedback characteristics depending on their profile [37]. For instance, previous results indicate that "tensed" and "non-autonomous" people (based on the dimensions of the 16PF5 psychometric questionnaire [7]) are disadvantaged when controlling BCIs [16]. Interestingly, "non-autonomous" people are persons who rather learn in a social context [7]. "Tensed" people might also benefit from a reassuring social presence and emotional feedback.

In a previous BCI experiment, we analysed the influence of a learning companion, i.e., a type of educational agents which can provide a complex form of social presence and emotional feedback in a controlled environment. During this last experiment, we designed, implemented and tested the first artificial learning companion dedicated to BCI user training [38]. This learning companion was called PEANUT for Personalized Emotional Agent for Neurotechnology User Training (see Figure 1). In between two trials, PEANUT provided the learners with social presence and emotional feedback through interventions that were composed of both spoken sentences and displayed facial expressions. The interventions were selected based on the current and previous performances of the learner. We found that such learning companion had a differential impact on the participants' performances depending on their autonomy. Also, the presence of a learning companion influenced how the participants felt about their ability to learn and memorize how to use a BCI, which is a dimension of the user experience that we assessed. Thus, we found that a learning companion providing a complex form of social presence and emotional feedback could influence BCI user training outcomes, i.e., performances and user-experience.

1.2 Role of experimenters

Very little is known regarding the most prevalent and complex source of social presence and emotional feedback during experiments which originates from the human supervision (e.g., experimenter or caregiver). In experimental settings, experimenters present BCIs to the learners, ensure the smooth progress of the experiment and might also have an influence on users'

Figure 1: A participant training to perform mental tasks on the right with PEANUT, the first learning companion dedicated to MI-BCI user training, on the left.

states. For instance, in a clinical study, Hammer et al. report "we tried to keep the subjects motivated and attentive by providing non-alcoholic beverages, sweets and fresh air" [11]. It has been shown that users' states (e.g., motivation, attention) can influence the accuracy of MI-BCI classification [11]. However, the influence that experimenters might have on users' states and BCI training outcomes remains unknown and was not formally investigated. Only very few studies in clinical BCI-based motor rehabilitation post-stroke acknowledge and explicit the role of the therapists, without formally assessing their influence [29, 36, 46]. Rosenthal, who was part of the first in social psychology to stress the importance of studying the influence of experimenters, describes experimenters as "imperfect tools" [44]. Indeed, the literature from different fields states that experimenters may consciously or unconsciously affect their results. Experimenters can influence participants' responses, behaviour and performances via direct and/or indirect interactions [45]. There are several types of possible experimenter-related influence, one of them being psychosocial factors. Stereotyped people tend to behave in a stereotype-consistent way [53]. For example, elderly people tend to walk more slowly or to have impaired memory performances if they feel stereotyped [53]. The "experimenter demand effect" is another example of experimenter-related influence. It can occur when participants unconsciously try to fit the appropriate image reflected by the experimenter's behaviour and therefore want to please and assist the experimenters in obtaining their expected results [44]. These different influences can be modulated through the experimenters' own characteristics (e.g., gender, age, ethnicity and professional status) and/or

behaviour (e.g., gaze, touch and verbal interactions) [44].

Among experimenters' characteristics modulating their influence, one of the most prevalent seems to be the gender. Previous experiments often report a simple effect of the experimenters' gender or an interaction between experimenters' and participants' gender on experimental outcomes [21, 44, 48]. Indeed, many cultural stereotypes are gender-based. One of which is that women have weaker math abilities than men. In previous experiments, Spencer et al. found that depending on women being told that difficult maths tests were respectively gender-dependent or independent, they did underperform or not compared to men participants [48]. In the neurofeedback field where users are trained to control their brain activity, Wood and Kober found that experimenters could have a differential impact on neurofeedback training depending on three parameters : experimenters' gender, participants' gender and participants' level of locus of control in dealing with new technologies [54]. They relate this difference of performances to psychosocial factors.

An interaction between the experimenter's and the participant's gender can also modulate the experimenter demand effect. For instance, when participants are instructed by an experimenter from the opposite sex, they seem more likely to act in ways that confirm the experimenter's hypothesis [32]. Also, neurophysiological responses associated with defensiveness, i.e., the aim to avoid being criticised, is associated with greater relative left frontal activation in the presence of experimenters from the opposite sex compared to experimenters from the same sex [18]. Thus, an interaction of experimenters' and participants' gender can influence experimental outcomes, including neurological responses measured using EEG [8, 18].

1.3 Research hypotheses

Literature in the field has identified direct factors that affect user learning (e.g., motivation, attention), although there influence is still understudied. In order to improve BCI reliability, it is thus highly relevant to identify, control and manipulate the factors affecting users' states. Among these many factors (e.g., instructions, feedback or exercise design) our literature review presented above suggests that the experimental environment may have a major influence, notably experimenters [42]. Despite the central role that experimenters have in BCI experimental process and the literature regarding the impact of social presence and emotional feedback, no studies had yet been led to evaluate their influence on MI-BCI experimental

outcomes, i.e., performances and user-training.

Experimenter's profile includes many aspects such as age or personality. As described in Section 1.2, literature from other fields suggests that one of the most prevalent characteristics modulating experimenters' influence seems to be the gender. Indeed, experimental outcomes (including neurological responses) may be significantly influenced by gender-related factors. Such impact might differ depending on the profile of the participants and experimenters. Therefore, based on the literature, we formulated the following hypotheses:

- (**H1 - MI-BCI performances**) MI-BCI performances undergo a gender-related influence of experimenters, possibly modulated by users' gender.
- (**H2 - User experience**) User experience undergo a gender-related influence of experimenters, possibly modulated by users' gender.
- (**H3 - Experimenters' and participants' profile**) These effects are modulated by experimenters' and participants' profile.

The remainder of this paper is organized as follows. In Section 2 -Materials & methods-, we provide information regarding the implementation of the experimental protocol that enabled us to test these hypotheses. Then, in Section 3 -Results- and in Section 4 -Discussion-, we respectively report and discuss the results from our experiment[1]. Finally, in Section 5 -Conclusions and Prospects-, we offer a conclusion on the matter as well as ideas and recommendations for future research.

2 Materials & methods

2.1 Participants

Sixty healthy MI-BCI naïve participants (29 women ; age 19-59, $M = 29$, $SD = 9.32$) completed the study. None of them reported a history of neurological or psychiatric disorder. Six

[1]Preliminary results regarding the interaction of experimenters' and participants' gender on the evolution of MI-BCI performances were previously published in a short conference paper presented at the 8th International BCI Conference [43]. Here we present additional and more complete results regarding potential confounding factors such as motor-related artefacts. We also present for the first time results related to the participants' psychological profile that provide first leads toward a better understanding of this experimenters' influence. Finally, we also provide new user-experience related results.

experimenters conducted the study (3 women ; age 23-37, $M = 29.2$, $SD = 5.6$) among whom two (1 woman) were experienced in BCI experimentation, having conducted more than 100 hours of EEG-based BCI experiments, and four were beginners (2 women) who were trained to perform a BCI experiment beforehand. All beginner experimenters were trained in a reproducible way by the experienced experimenters. Each experimenter was randomly assigned to 10 participants (5 women and 5 men) that they had never met before the session. All experimenters had the same ethnicity, i.e., Caucasian white native french, and were asked to wear their usual work clothing (casual, not extravagant, not sexualized). This choice was made in order to investigate the potential influence of experimenters in usual BCI experimental settings.

Our study was conducted in accordance with the relevant guidelines for ethical research according to the Declaration of Helsinki. Both participants and experimenters gave informed consent before participating in the study. In order to avoid biased behaviour, this study was conducted using a deception strategy, partially masking the purpose of the study. Participants were told that the study aimed at understanding which factors (unspecified) could influence BCI outcomes, i.e., performances and/or user experience. Experimenters were aware of the goal of the study. The study has been reviewed and approved by Inria's ethics committee, the COERLE (Approval number: 2018-13).

2.2 Experimental protocol

Each participant completed one session of 2 hours with a MI-BCI. During this session, participants were first asked to read and sign the consent form and complete several questionnaires (see the following Subsection 2.3 -Questionnaires-), which took around 20 min. Once the EEG cap (see Subsection 2.4 -EEG Recordings & Signal Processing-) was placed on their head, the participants performed six 7-minutes runs during which they had to learn to perform two MI tasks with the BCI, i.e., imagine right or left hand movements (around 60 min, including breaks between the runs). Finally, the participants were asked to fill the post-session questionnaires, the EEG cap was uninstalled and a debriefing was made (around 15 min).

The Graz training protocol was used [35]. It is divided into two steps: first, the training of the system and second, the training of the user. The first two runs were used as calibration in order to provide to the system examples of EEG patterns associated with each of the MI tasks.

During the first two runs, as the classifier was not yet trained to recognize the mental tasks being performed by the user, it could not provide a consistent feedback. In order to limit biases with the other runs, e.g., EEG changes due to visual processing differences between runs, the user was provided with an equivalent sham feedback, i.e., a blue bar randomly appearing and varying in length. These two steps and their respective runs are visually depicted in Figure 2.

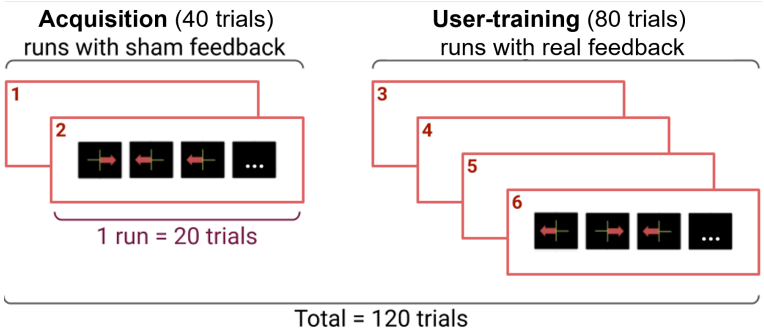

Figure 2: The BCI session included 6 runs divided into two steps: (1) data acquisition to train the system (2 runs) and (2) user training (4 runs). After Run 2, the classifier is trained on the data acquired during the two first runs.

During each run, participants had to perform 40 trials (20 per MI-task, presented in a random order), each trial lasted 8s. At t = 0s, a cross was displayed on the screen. At t = 2s, an acoustic signal announced the appearance of a red arrow, which appeared one second later (at t = 3s) and remained displayed for 1.25s. The arrow pointed in the direction of the task to be performed, namely left or right to imagine a movement of the left or right hand. Participants are instructed to start performing the corresponding MI-task as soon as the arrow appeared, and to keep doing so until the cross disappeared. Finally, from t = 4.25s, a visual feedback was continuously provided in the shape of a blue bar, the length of which varied according to the BCI classifier output. Only positive feedback was displayed, i.e., the feedback was provided only when the instruction matched the recognized task. The feedback was provided for 3.75s and was updated at 16Hz, using a 1s sliding window. After 8 seconds, the screen turned black again until the beginning of the next trial. The participant could then rest for a few seconds. The timeline of a trial is shown in Figure 3.

Following the recommendations from the literature, the participants were encouraged to perform a kinesthetic imagination [31] and to choose their own mental imagery strategies [19], e.g., imagining waving at someone or playing the piano. Participants were instructed to find a strategy for each MI task so that the system would display the longest possible feedback bar.

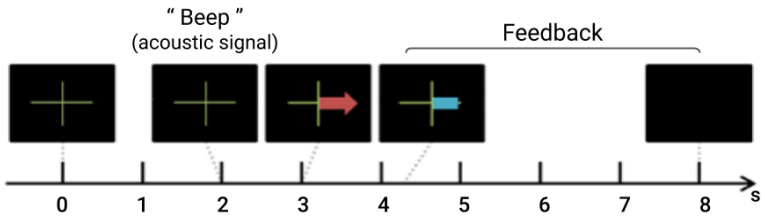

Figure 3: Timeline of a trial.

Instructions were written in advance and read by the experimenters so that all the participants started with the same standardized information. As they would have been in any standard BCI experiment, the experimenters were free to interact with the participants before, during and after the experiment, e.g., seating and/or standing. They were in charge of welcoming participants in the lab, showing them the way to the experimental room, making them sign the consent form, explaining them what would happen during the whole experiment, setting up the EEG cap on them, asking them to fill-in various questionnaires, calibrating the BCI system and making it run for the participants, answering questions that the participants may have, providing them with water if they required some, removing the cap and debriefing with the participant at the end of the experiment. Experimenters were only asked not to reveal the aim of the experiment before its very end.

2.3 Questionnaires

As stated in the introduction, the personality and the cognitive profile of participants and experimenters can respectively influence BCI performances and the experimenter bias [37]. Therefore, we assessed the personality and the cognitive profile of both the participants and the experimenters. The 5th edition of the 16 Personality Factors (16PF5), i.e., a validated psychometric questionnaire to assess different aspects of people's personality and cognitive profile was filled by both experimenters and participants [7]. This questionnaire identifies 16 primary factors of personality, including tension and autonomy. Participants also completed a mental rotation test measuring their spatial abilities [51].

The participants also filled pre and post experiment questionnaires especially developed for BCI purpose by Hakoun et al. These questionnaires assessed the participants' states and the user-experience [3, 13]. Based on validated questionnaires, it determines five dimensions of the user-state and/or the user-experience. Three dimensions, i.e., the mood, mindfulness and

motivational states, were assessed pre and post training. The evolution of the participant's states provides an information regarding the user-experience. Two dimensions, i.e., the cognitive load (amount of cognitive process required to control the MI-BCI system) and the sense of agency (feeling of control of the participant over the feedback provided by the MI-BCI) assessed the user-experience post-training.

2.4 EEG Recordings & Signal Processing

To record the EEG signals, 27 active scalp electrodes, referenced to the left earlobe, were used (Fz, FCz, Cz, CPz, Pz, C1, C3, C5, C2, C4, C6, F4, FC2, FC4, FC6, CP2, CP4, CP6, P4, F3, FC1, FC3, FC5, CP1, CP3, CP5, P3, 10-20 system). The electromyographic (EMG) activity of the hands was recorded using two active electrodes located 2.5cm below the skinfold on each wrists. The electrooculographic (EOG) activity of one eye was recorded using three active electrodes. Two of them were located below and above the eye and one was located on the side. They aimed at recording vertical and horizontal movements of the eye. Physiological signals were measured using a g.USBAmp (g.tec, Austria), sampled at 256 Hz, and processed online using OpenViBE 2.1.0 [41]. To classify the two MI tasks from EEG data, we used participant-specific spectral and spatial filters. To do so, we used the now standard algorithms proposed by Blankertz et al. in [5]. More precisely, from the EEG signals recorded during the calibration runs, we first identified a participant-specific discriminant frequency band using the heuristic algorithm proposed in [5] (Algorithm 1 of that paper). Roughly, this algorithm selects the frequency band whose power in the sensorimotor channels maximally correlates with the class labels. Here we used channels C3 & C4 after spatial filtering with a Laplacian filter as sensorimotor channels, as recommended in [5]. The algorithm selected a discriminant frequency band within the interval from 5 Hz to 35 Hz, with 0.5Hz large bins. Once this discriminant frequency band automatically identified, we filtered EEG signals in that band using a Butterworth filter of order 5.

Then, still has recommended in [5], we used the Common Spatial Pattern (CSP) algorithm [40], in order to optimize 3 pairs of spatial filters, still using the data from the two calibration runs. Such spatially filtered EEG signals should thus have a band power which is maximally different between the two MI conditions. We then computed the band power of these spatially filtered signals by squaring the EEG signals, averaging them over a 1 second sliding window

(with 1/16th second between consecutive windows), and log-transforming the results. This led to 6 different features per time window, which were used as input to a Linear Discriminant Analysis (LDA) classifier [24]. As mentioned above, this LDA was calibrated on the data from the two calibration runs. These filters and classifier were then applied on the subsequent runs to provide online feedback. It should be noted that this BCI design and EEG signal processing is a rather standard approach, that has been used in numerous previous experiments by various laboratories, see, e.g., [4, 5, 14].

2.5 Variables, Factors & Statistical analyses

As presented in the introduction, our experiment aimed at testing three different hypothesis. In the following Subsections we present the variables, factors and statistical analyses used to test each of these hypotheses. The statistical analyses mostly consist of ANOVAs, that are considered as robust against the normality assumption. To the best of our knowledge, no other non parametric test enabled to perform the analysis that we were interested in. Spearman or Pearson correlations were also obtained depending on the distribution of the data collected (assessed using Shapiro-Wilk tests).

2.5.1 H1 - MI-BCI performances

To test our first hypothesis (H1), i.e., MI-BCI performances undergo a gender-related influence of experimenters, possibly modulated by users' gender, two measures of performance were used.

The first performance metric we used is the online Trial-wise Accuracy (TAcc). This metric is computed by first summing the (signed) LDA classifier outputs (distance to the separating hyperplane) over all epochs (1s long epochs, with 15/16 s overlap between consecutive windows) during the trial feedback period. If this sum sign matched the required trial label, i.e., negative for left hand MI and positive for right hand MI, then the trial was considered as correctly classified, otherwise it was not. The TAcc for each run was estimated as the percentage of trials considered as correctly classified using this approach. TAcc is the default accuracy measure provided online in the MI-BCI scenarios of OpenViBE, and the only performance metric that the experimenters were seeing online. It should be noted that this metric takes into account the classifier output and is thus also related to the feedback bar length as it is propor-

tional to the classifier output. Our participants were instructed to train to obtain not only a correct classification, but also a feedback bar as long as possible, the TAcc metrics thus take into account both aspects. Offline, we also computed the Epoch-wise Accuracy (EAcc) as the percentage of epochs (1s long time windows) from the feedback periods that were correctly classified. Thus, this metric only considers whether the classification was correct, but not the feedback bar length as it does not take into account the classifier output. However, it does reflect how often EEG epochs were correctly classified, and thus how often the subjects received correct positive feedback. It is also a rather standard classification performance metrics in BCI Machine Learning [49], we thus also provide it for reference.

These two measures of MI-BCI performances over the series of 4 user-training runs, i.e., *"Run"*, were then used in two 3-way repeated measures mixed ANOVAs with *"ExpGender"*, *"ParGender"* and *"Run"* as independent variables and the repeated measures of performance over the runs, i.e., TAcc or EAcc, as dependent variable. The results are reported in Subsection 3.1 -H1 - MI-BCI performances-.

2.5.2 H2 - User experience

Second, we wanted to assess the potential impact of the experimenters' and participants' gender on the user experience (H2). The user experience is defined by the two percentages provided by the questionnaire of Hakoun et al. [3, 13] regarding the amount of cognitive load and sense of agency felt during the training. It is also defined by the evolution of mood, mindfulness and motivation of the participants between the beginning and end of the training. This evolution is assessed by subtracting the measure post training to the measure pre training, both assessed in percent. The higher the percentages are, the more participants increased their reported levels of positive emotions and calm, mindfulness, motivation, cognitive load and sense of agency.

These five measures of user experience were then used in five 2-way ANOVAs or ANCOVAs, one per dimension, with *"ExpGender"* and *"ParGender"* as independent variables and either the measure of cognitive load, sense of agency, mood, mindfulness or motivation as dependent variable. Performances averaged over all runs, i.e., TAcc or EAcc, were used as covariate if they were correlated to the dependent variable. The results are reported in Subsection 3.2 -H2 - User experience-.

2.5.3 H3 - Experimenters' and participants' profile

Finally, we wanted to know if other characteristics of the experimenters' and/or participants' profile than the gender could provide first elements of comprehension regarding the potential difference in MI-BCI performances or user-experience (H3). We focused on characteristics of the profile that were shown to have an influence on BCI performances in previous studies, i.e., mental rotation scores (MRS), tension and autonomy [16]. Participants with low MRS [51], tensed and/or non-autonomous (both measured using the 16PF5 questionnaire [7]) were shown to have lower BCI performances than the others [16].

The groups formed by experimenters' and participants' gender did not have similar MRS and autonomy. Thus, we assessed the influence of these two measures on the results obtained for H1 using the same ANOVAs that were used to test the hypothesis (two 3-way repeated measures mixed ANOVAs with *"ExpGender"*, *"ParGender"* and *"Run"* as independent variables and the repeated measures of performance over the runs, i.e., TAcc or EAcc, as dependent variable) and the autonomy, i.e., *"Autonomy"*, or the mental rotation score, i.e., *"MRS"*, of the participants as covariate. The results are reported in Annex B -Details regarding the analyses on the potential influence of MRS and autonomy differences in participant groups-.

Then, we focused on the potential influence of the tension. We separated the participants into two groups depending on their tension *"ParTension"*. The threshold between high and low tension was defined using the median tension score (i.e., median of 6, low and high tension respectively corresponding to scores of [1, 5] and [6, 10], 10 being the maximum). We performed two 3-way ANOVAs with *"ParTension"*, *"ExpGender"* and *"ParGender"* as independent variables and one of the measures of performance averaged over all runs, i.e., TAcc or EAcc, as dependent variable. The results are reported in Subsection 3.3 -H3 - Experimenters' and participants' profile-.

3 Results

Among the 60 participants, 1 participant did not complete all of the four runs of participant training due to a technical issue and 3 outperformed the others (by more than two SDs) both in term of TAcc (respectively, outliers $Ms1 = 98.13$, $Ms2 = 98.13$, $Ms3 = 99.38$; $Mgrp = 62.78\%$, $SDgrp = 16.2$) and EAcc (outliers $Ms1 = 88.94$, $Ms2 = 90.36$, $Ms3 = 94.51$; $Mgrp = 59.33\%$,

SDgrp = 12.3). Thus, the following analyses are based on the results of 56 participants (27 women).

The automatically selected and subject-specific discriminant frequency bands used to classify the two MI tasks from EEG data were in the range of 16.4 $\pm$ 7.78 Hz to 19.58 $\pm$ 7.44 Hz with an average length of 3.17 $\pm$ 2.99 Hz (see Subsection 2.4 -EEG Recordings & Signal Processing-).

Before it all, we verified that groups formed by participants' gender, i.e., *"ParGender"*, and experimenters' gender, i.e., *"ExpGender"*, had comparable profiles. To check that groups were comparable, we ran 2-way ANOVAs with *"ExpGender"* and *"ParGender"* as independent variables and either MRS, tension or autonomy as dependent variable.

Results indicate that groups are comparable in terms of tension. Though, participants' gender influence their MRS [$F(1, 52) = 17.47$, $p \leq 10^{-3}$, $\eta^2 = .25$]. Men (*Mmen* = 0.07, *SD* = 0.02) had higher MRS than women (*Mwomen* = 0.05, *SD* = 0.02), which is in accordance with the literature [22]. Furthermore, participants training with men or women experimenters did not have the same level of autonomy [$F(1, 52) = 4.01$, $p = .05$, $\eta^2 = .07$]. Participants training with men experimenters (*MmenExp* = 6.35, *SD* = 1.74) were more autonomous than participants training with women experimenters (*MwomenExp* = 5.67, *SD* = 1.66). As the autonomy and MRS of participants was found to influence their BCI performances [16], we controlled for the potential influence of these variables in our subsequent analyses (see Appendix B - Details regarding the analyses on the potential influence of MRS and autonomy differences in participant groups-).

In the following sections, we report the results of the analyses presented in Section 2.5 that we performed to test each of our hypotheses.

3.1 H1 - MI-BCI performances

We started by testing the H1 hypothesis, i.e., MI-BCI performances undergo a gender-related influence of experimenters, possibly modulated by users' gender. As stated in 2.5.1 -H1 - MI-BCI performances-, we performed two 3-way repeated measures mixed ANOVAs with *"ExpGender"*, *"ParGender"* and *"Run"* as independent variables and the repeated measures of performance over the runs, i.e., TAcc or EAcc, as dependent variable.

First, we performed such ANOVA using the TAcc. After correction of sphericity using the

Huynh-Feldt method ($\epsilon = 0.92$), the results revealed no simple effect of *"Run"* [$F(2.8, 144) = 1.81$, $p = .15$, $\eta^2 = .03$], *"ExpGender"* [$F(1, 52) = 0.54$, $p = .47$, $\eta^2 = .01$] nor *"ParGender"* [$F(1, 52) = 0.09$, $p = .76$, $\eta^2 = .01$]. They also revealed no interaction of *"Run*ExpGender"* [$F(2.8, 144) = 0.08$, $p = .96$, $\eta^2 = 10^{-2}$] nor *"ParGender*ExpGender"* [$F(1,52) = 0.60$, $p = .44$, $\eta^2 = .01$]. Though, the ***"Run*ParGender"* interaction was significant [$F(2.8, 144) = 5.98$, $p = .001$, $\eta^2 = .1$]**. Figure 4 represents the evolution of the participants' TAcc depending on their gender.

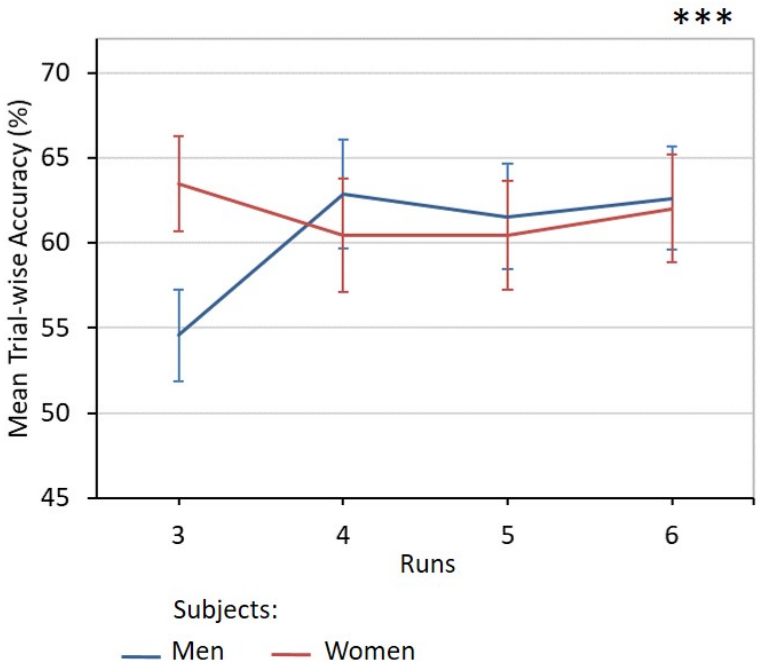

Figure 4: TAcc evolution depending on participants' gender.

A significant ***"Run*ParGender*ExpGender"* interaction was found as well [$F(2.8, 144) = 3.46$, $p = .02$, $\eta^2 = .06$]**. Figure 5 represents the participants' TAcc evolution depending on the experimenters' and participants' gender.

Next, we performed this same analysis using the EAcc. After correction of sphericity using the Huynh-Feldt method ($\epsilon = 0.8$), the results revealed no simple effect of *"Run"* [$F(2.4, 125) = 1.53$, $p = .22$, $\eta^2 = .03$], *"ExpGender"* [$F(1, 52) = 0.26$, $p = .61$, $\eta^2 \leq 0.01$] and *"ParGender"* [$F(1, 52) = 0.23$, $p = .64$, $\eta^2 \leq 0.01$]. They revealed no interaction of *"Run*ParGender"* [$F(2.4, 125) = 1.92$, $p = .14$, $\eta^2 = .04$], *"Run* ExpGender"* [$F(2.4, 125) = 0.23$, $p = .83$, $\eta^2 = 0.01$] nor *"ParGender*ExpGender"* [$F(1, 52) = 0.92$, $p = .34$, $\eta^2 = .02$]. Finally, the interaction of *"Run*ParGender*ExpGender"* [$F(2.4, 125) = 1.38$, $p = .26$, $\eta^2 = .03$] was not

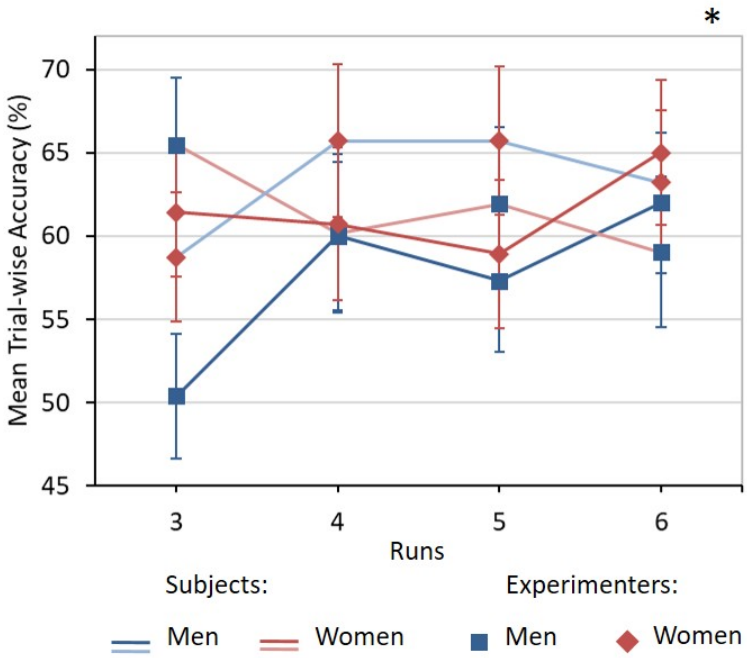

Figure 5: TAcc evolution depending on the experimenters' and participants' gender.

significant either.

We controlled for the potential influence of the most common artefact sources, i.e., electrooculography (EOG) and electromyography (EMG) [10], on our performances measures, i.e., TAcc and EAcc, in additional analyses that are presented in Appendix A -Details regarding the analyses on the potential influence of artefact sources-. These analyses did not reveal an influence of EOG or EMG artefacts that could have affected the EEG-based BCI performances.

We also controlled for the potential influence of MRS and autonomy differences in participant groups formed using the participants' and experimenters' gender. These analyses are presented in Appendix B -Details regarding the analyses on the potential influence of MRS and autonomy differences in participant groups- and did not reveal any potential bias from MRS and autonomy differences in participant groups.

3.2 H2 - User experience

Then, we tested the H2 hypothesis, i.e., user experience undergo a gender-related influence of experimenters, possibly modulated by users' gender. As stated in 2.5.1 -H1 - MI-BCI performances-, we analysed the influence of participants' and experimenters' gender on five

indicators of user-experience, i.e., mood, mindfulness, motivation, cognitive load and sense of agency.

First, we checked if the performances had an impact on the reported user-experience measures. We found that the sense of agency post training was positively correlated to both the TAcc [Spearman correlation, r(56) = .38, $p < 10^{-2}$] and EAcc [Spearman correlation, r(56) = .34, p = .01] metrics.

Then, we performed five 2-way ANOVAs or ANCOVAs, one per dimension, with *"ExpGender"* and *"ParGender"* as independent variables and either the measure of cognitive load, sense of agency, mood, mindfulness or motivation as dependent variable. Performances averaged over all runs, i.e., TAcc or EAcc, were used as covariate if they were correlated to the dependent variable.

We did not find any significant single effect or interaction including the experimenters' gender for the cognitive load, sense of agency, mood, mindfulness or motivation (see Appendix C).

We only found a **significant influence of *"ParGender"* [$F(1, 52) = 6.23$, $p = .02$, $\eta^2 = .11$] on the difference of mindfulness post and pre training**. Overall, men participants had a decrease of mindfulness (*MmindfulnessMen* = -8.33 $\pm$ 3.01) whereas women participants had an increase of mindfulness (*MmindfulnessWomen* = 2.5 $\pm$ 3.12) over the session.

3.3 H3 - Experimenters' and participants' profile

As presented in the introduction, a previous study has shown that participants' autonomy and tension both respectively correlate positively and negatively with BCI performances [16]. As there were differences in autonomy between the participant groups formed by experimenters' and participants' gender, we analysed the potential influence of participants' autonomy in specific analyses whose results are presented in Appendix B -Details regarding the analyses on the potential influence of MRS and autonomy differences in participant groups-. These analyses did not reveal any potential influence of the differences in autonomy and MRS on our results. Thus, in this Section, we only focused on the tension to perform analyses related to the psychological profile of the participants and experimenters. High tension scores computed from the 16PF5 questionnaire indicate highly tensed, impatient and frustrated personalities whereas low scores indicate relaxed, patient and composed personalities.

3.3.1 Assessing the influence of participants' tension

We checked if an influence of participants' tension could be found in our results by performing an analysis of correlation between participants' tension and our measures of performance. It revealed a **negative correlation between participants' tension and both the TAcc [Spearman correlation, r(56) = -.39, $p < 10^{-2}$] and EAcc [Spearman correlation, r(56) = -.29, p = .03] metrics**, which is in accordance with previous results [16].

Therefore, we investigated if the tension could explain the differences of performances' depending on the participants' and experimenters' gender. As stated in 2.5.1 -H1 - MI-BCI performances-, we performed two 3-way ANOVAs with *"ParTension"*, *"ExpGender"* and *"ParGender"* as independent variables and one of the measures of performance averaged over all runs, i.e., TAcc or EAcc, as dependent variable.

When using the TAcc as a measure of performance, we did not find any simple effect of *"ExpGender"* [$F(1, 48) = 1.51$, $p = .23$, $\eta^2 = .03$], nor *"ParGender"* [$F(1, 48) = 1.72$, $p = .2$, $\eta^2 = .04$]. Though, a trend toward a weak impact of *"ParTension"* was found [$F(1, 48) = 3.8$, $p = .06$, $\eta^2 = .07$]. No interactions were found for *"ExpGender*ParGender"* [$F(1,48) < 10^{-3}$, $p = 1$, $\eta^2 < 10^{-3}$], *"ParTension*ParGender"* [$F(1, 48) = 0.18$, $p = .67$, $\eta^2 < 10^{-2}$], *"ParTension*ExpGender*ParGender"* [$F(1, 48) = 0.47$, $p = .5$, $\eta^2 = .01$]. Though a **significant and strong interaction was found between *"ParTension*ExpGender"* [$F(1, 48) = 18.94$, $p < 10^{-3}$, $\eta^2 = .28$]**.

When using the EAcc as measure of performance we did not find any simple effect of *"ExpGender"* [$F(1, 48) = 1.12$, $p = .3$, $\eta^2 = .02$], nor *"ParGender"* [$F(1, 48) = 2.59$, $p = .11$, $\eta^2 = .05$]. Though, a weak but significant impact of *"ParTension"* was found [$F(1, 48) = 4.43$, $p = .04$, $\eta^2 = .08$]. No interactions were found for *"ExpGender*ParGender"* [$F(1, 48) = 0.02$, $p = .89$, $\eta^2 < 10^{-3}$], *"ParTension*ParGender"* [$F(1, 48) = 0.1$, $p = .75$, $\eta^2 < 10^{-2}$], *"ParTension*ExpGender* ParGender"* [$F(1, 48) = 0.72$, $p = .1$, $\eta^2 = .02$]. Though, a **significant interaction was found between *"ParTension*ExpGender"* [$F(1, 48) = 21.98$, $p < 10^{-3}$, $\eta^2 = .31$]**.

Figure 6 represents the average performances of participants with tensed and non-tensed personalities when taking into account the gender of their experimenters. Non-tensed participants seem to have higher performances, i.e., TAcc and EAcc, when training with women experimenters while tensed participants seem to have higher performances, i.e., TAcc and EAcc,

when training with men experimenters.

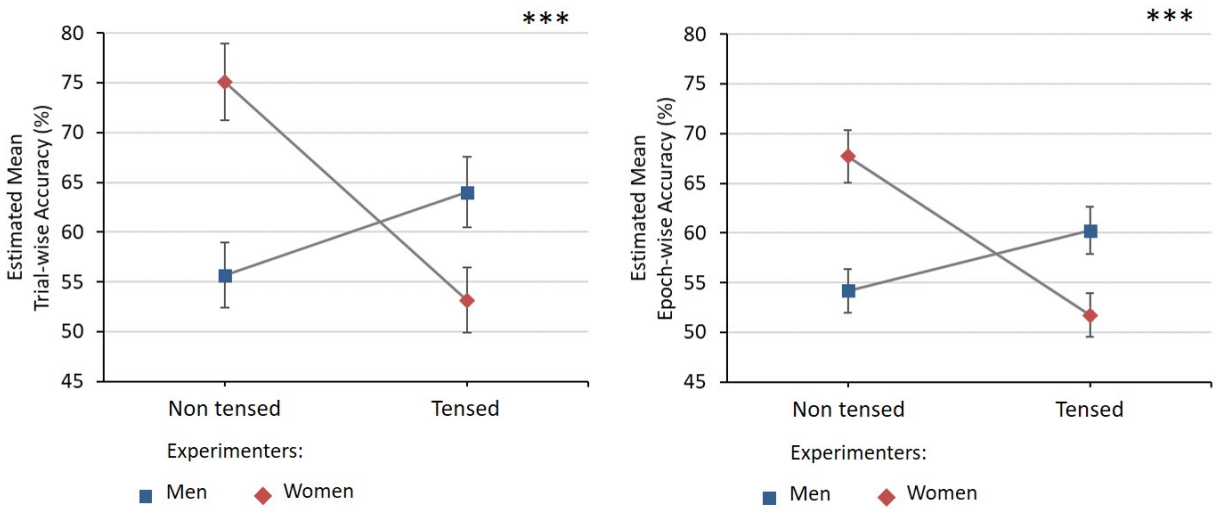

Figure 6: Estimated mean performances depending on participants' tension and experimenters' gender.

3.3.2 Assessing the influence of experimenters' tension

Previous results found that a similarity between participants' and experimenters' profile could lead to higher bias in experimental results [44]. As participants' level of tension had a significant impact on their results, we analysed the potential influence of the level of tension of our experimenters. The tension score in the personality of the three men and three women experimenters were respectively of [5, 5 and 7] and [3, 4 and 5], indicating a higher level of tension among men experimenters than among women experimenters. Therefore, we investigated further to know if the influence of the experimenters' came from a psychosocial factor related to their gender or from their level of tension which was higher among men experimenters than women participants.

We checked if, independently of gender, there was of correlation between the tension of the experimenter and the performances of the participants. We did not find any correlation of the experimenters' tension with the TAcc [Spearman correlation, r(56) = .03, p = .83], nor with the EAcc [Spearman correlation, r(56) = .11, p = .44].

3.4 Summary of the results

Hypothesis	Analyses	Significant results
H1- MI-BCI performances	3-way repeated measures mixed ANOVA with "ExpGender", "ParGender" and "Run" as independent variables and the repeated measures of TAcc performance over the runs as dependent variable	*"Run*ParGender"* [$F(2.8, 144) = 5.98$, $p = .001$, $\eta^2 = .1$] *"Run*ParGender*ExpGender"* [$F(2.8, 144) = 3.46$, $p = .02$, $\eta^2 = .06$]
H2- User experience	2-way ANOVA with *"ExpGender* ParGender"* as independent variables and the measure of mindfulness as dependent variable	*"ParGender"* [$F(1, 52) = 6.23$, $p = .02$, $\eta^2 = .11$]
H3- Experimenters' and participants' profile	Spearman correlation	Negative correlation between participants' tension and both TAcc [Spearman correlation, r(56) = -.39, $p < 10^{-2}$] and EAcc [Spearman correlation, r(56) = -.29, $p = .03$]
	3-way ANOVA with *"ParTension"*, *"ExpGender"*, *"ParGender"* as independent variables and the TAcc measures of performance averaged over all runs as dependent variable	*"ParTension*ExpGender"* [$F(1, 48) = 18.94$, $p < 10^{-3}$, $\eta^2 = .28$]
	3-way ANOVA with *"ParTension* ExpGender*ParGender"* as independent variables and the EAcc measures of performance averaged over all runs as dependent variable	*"ParTension"* [$F(1, 48) = 4.43$, $p = .04$, $\eta^2 = .08$] *"ParTension*ExpGender"* [$F(1, 48) = 21.98$, $p < 10^{-3}$, $\eta^2 = .31$]

Table 1: Summary of the significant results per hypothesis.

4 Discussion

In the following Subsections, we discuss the results obtained for each of our hypothesis.

4.1 H1 - MI-BCI performances

To test the H1 hypothesis, i.e., MI-BCI performances undergo a gender-related influence of experimenters, possibly modulated by users' gender, we used two metrics of performances. The TAcc, which represented what the participants were instructed to improve during training, and the EAcc, a traditional measure of BCI performances. We did not find a single influence of the experimenters' and/or participants' gender on these performances. Though, we found a significantly different evolution across runs of the TAcc between men and women participants (see Figure 4). Women participants seemed to start the training with already good TAcc, which decreased during the second run and increased again during the last run. Men participants, however, started with rather low TAcc and then drastically improved during the second run and then stagnated to reach slightly higher final TAcc than women.

In addition, experimenters' gender seemed to have an influence on this previous interaction. Indeed, the evolution of the TAcc appears to depend on participants' and experimenters' gender (see Figure 5). We found the same tendency for men participants to start with lower TAcc at the beginning of the session independently of the experimenter's gender. However, men seemed to start with drastically lower TAcc when they were training with men experimenters. They also seemed to have higher TAcc throughout the session when they were training with women experimenters. Women participants seemed to start with higher TAcc when training with men experimenters, though their TAcc tended to drop throughout the session. However, when training with women experimenters, they seemed to have a great increase in TAcc during the last run. In social psychology, Nichols and Maner found that participants who are instructed by an opposite-sex experimenter tend to confirm the experimenter's expectation regarding the experimental results [32]. The initial performances (during R3) are consistent with their findings. However, this does not seem to hold true for the evolution of the participants' performances.

Interestingly enough, our results regarding the impact of participants' and experimenters' gender do not match those of a recently published neurofeedback study [54]. We do concur on the fact that an interaction of participants' and experimenters' gender influences performances. Though, Wood and Kober found that the combination of women participants training with women experimenters hampered the training outcomes of the participants [54]. They observed no learning effect in this group. The influence of the participants' tension found in our results might partly explain this difference of results. In their article, they found a strong and significant positive correlation between the locus of control in dealing with technology, i.e., the level of control that people feel that they have over the control of a technology, and the performances of women participants training with women experimenters. We did not assess this trait of our participants, thus the difference in results might also arise from a difference in the locus of control of our women participants. Even though the locus of control of our participants was not assessed, we assessed the sense of agency they felt toward the feedback that their were provided with during the training. We did not observe any gender influence over the sense of agency reported by our participants. Overall, our analysis of the user-experience metrics only revealed an influence of participants' gender on the evolution of the mindfulness metric. Men participants tended to have a decrease of mindfulness over the session, when

women participants tended to increase their level of mindfulness. Also, Wood and Kober do not report controlling for the prior acquaintanceship between their participants and experimenters [54]. Rosenthal found that this could modulate the bias induced by experimenters mostly between men experimenters and women participants [44]. Another explanation of the differences found between our two studies would be that, as stated by Wood and Kober, by asking their participants to fill a questionnaire regarding their locus of control in dealing with technology, they might have activated a stereotype bias [54]. Such stereotype was not activated in our study. Finally, the protocol used by Wood and Kober was a neurofeedback one aiming at up-regulating the sensorimotor rhythm, and not a two-commands MI-BCI training. This most possibly also contributes to the differences of results obtained.
Current results do not seem to be biased by the mental rotation scores nor the autonomy of the participants. Indeed, the same analysis that led us to these conclusions were run with these variables as covariate. Results do not reveal any impact of these variables, and revealed the same significant effect as mentioned above. Artefacts potentially arising from eye or hand movements did not seem to bias of our results either.

4.2 H2 - User experience

Our results did not indicate any influence related to the gender of the experimenter on the participants' user experience. Such influence could have been expected based on previous results indicating that a social presence and an emotional feedback provided through the use of a learning companion impacted one dimension of the user experience, i.e., how the participants felt about their ability to learn and memorize how to use a MI-BCI. Further experiments using different metrics of the user experience might provide more insight on the potential influence of experimenters on the user experience.

4.3 H3 - Experimenters' and participants' profile

When investigating the influence of the tension of the participants on these results, we found results that tend to be in accordance with the ones of Jeunet et al. [16]. Participants with tensed personality trait tend to have lower performances than non-tensed participants. An influence of participants anxiety was already found in early researches on regulation of alpha [50]. Our results revealed that the influence of the participants' tension on MI-BCI performances seems

to be modulated by the gender of the experimenter. Tensed and non tensed participants had better performances when training respectively with men experimenters and women experimenters. This result might provide a first lead toward understanding the interaction between experimenters' and participants' influence on MI-BCI performances. We did not find any significant influence of experimenters' tension on participants' performances. In the future, testing whether a similarity of experimenters' and participants' psychological profiles could lead to higher potential bias in the results would be of interest. In studies on social psychology, Rosenthal found that participants were more likely to respond to experimenters' expectancy when their level of anxiety was similar to their experimenter's level of anxiety [44]. He hypothesised that a similarity of experimenters' and participants' psychological profiles could lead to higher potential bias in the results. We can make the same hypothesis as Rosenthal to explain our results as men experimenters in our study had higher scores of tension than women experimenters [44]. Non-tensed participants might have been more inclined to respond to women experimenters' expectancy, i.e., to have high MI-BCI performances, who also tended not to be tensed. Tensed participants, however, might have been more inclined to respond to men experimenters' expectancy who also tended to be tensed. The number of participants did not enable to perform an analysis of both the experimenters' and participants' gender and tension at once, as the number of participants per group would have been too low. Furthermore, experimenters' level of tension was highly dependent on their gender. Larger scaled experiments with a greater number of experimenters would provide insight on this hypothesis.

4.4 Limitations

While this study does provide first insights on the interaction between experimenters' and participants' gender, future studies are needed to further explore it and explore its unknown long term influence. Studies with a larger number of experimenters and participants might provide more information regarding the underlying factors of this gender influence. For instance, it could confirm or disprove the interaction between experimenters' gender and participants' tension. If confirmed, our hypothesis regarding the beneficial similarity between the level of tension in participants' and experimenters' personality could be assessed.
Furthermore, our results might be explained by other factors. Indeed, inter-experimenter vari-

ability other than gender (e.g., teaching competence), intra-experimenter variability (e.g., appearance and outfit, fatigue, expectations), inter- and intra-participants variability (e.g., attractiveness, or motivation) - plus the interaction's characteristics (e.g., physical proximity, use of humour, familiarity, verbal and non-verbal communication, quantity of interaction, etc.) were not analysed. Indeed, many of these variables are very difficult to measure formally and objectively. Moreover, we were already measuring various aspects of the users and experimenters personality and states, using validated questionnaires, and the experiment was already long. Thus, measuring these additional factors would have required to remove some of the factors actually measured (to keep a reasonable experiment duration), which, according to the literature, were the one with the most influence, at least theoretically. In summary, our study shows an interaction between experimenters and participants on the evolution of MI-BCI performances. This interaction seems related to the experimenters' and participants' gender. However, future experiments should confirm and provide more insights regarding this interaction.

5 Conclusions and Prospects

In this paper, we investigated the presence of an experimenters' and participants' gender interaction on MI-BCI training outcomes, i.e., performances and user-experience. We led this work in response to the fact that previous BCI experiments indicated an influence of social presence and emotional feedback on BCI user training. Experimenters are the main source of such presence and feedback during BCI user training. Though, their impact on the MI-BCI user training outcomes remained unassessed. Also, results from different fields indicate that an interaction between experimenters' and participants' gender is likely to influence experimental outcome. Therefore, we asked 6 experimenters to each train 5 women and 5 men (60 participants in total) to perform right versus left hand motor imagery-BCI control over one session.

We did find an interaction between experimenters' and participants' gender on the evolution of trial-wise accuracy over a session. Furthermore, participants' mean performances were influenced by an interaction of the experimenters' gender and level of tension in participants' personality. No single effect or interaction related to the experimenters could be found on the

user-experience.

Our results highlight the need for research methods that formally take into account a greater amount of influencing factors (such as the experimenter) emerging from the experimental protocol and its context. For instance, the instructions that participants are provided with regarding the strategies they should adopt to perform mental-imagery tasks, are rarely formalized or mentioned in papers. Furthermore, most published experimental studies do not report taking into account the potential influence of experimenters. Both the literature and experimental results indicate that experimenter-related factors might explain part of the between-subject and/or between-study variability and contribute to the improvement and adaptation of MI-BCI training.

We argue that in the future the influence of experimenters should be considered carefully while designing and reporting experimental protocols. Such consideration would benefit many fields, in particular the Human Computer Interface and the BCI ones. A better understanding of the experimenters' influence could particularly lead to an improvement of MI-BCIs as they rely on a long and tedious user training during which experimenters have an important role. Other BCIs paradigms, such as P300-based BCIs[2], do not have such user training. However, regardless of the BCI paradigm or even field, during experimental studies assessing experimenter-unrelated factors and while experimenters' influence is not well understood, the bias that can arise from experimenters should be limited and controlled. Double-blind methods, in which neither the experimenters nor the participants know the group in which the participant is included, do limit the experimenter related bias. They are already used in clinical research. It would be worth applying similar methods in non-clinical experiments.

It should be noted that hiring research assistants to perform the experiments might not be a solution to limit experimenter-related bias. Indeed, it was shown that experimenters can unconsciously transmit their bias to their research assistants [44]. The literature suggests several other solutions to limit and control the potential bias arising from the experimenter [28, 45]. These methods include: monitoring participant-experimenter interaction, increasing the number and diversity of data collectors, pre-testing the method and controlling expectancy, providing an extensive training for administrators/ data collectors, monitoring and standard-

[2]P300-based BCIs rely on the elicitation of a characteristic neurophysiological response, i.e., the P300, following the presentation of an expected and unpredictable stimulus that the participants attend to.

izing the behaviour of experimenters with detailed protocol and pre-written instructions for the participant, and statistically controlling for bias. The use of learning companions, such as PEANUT (see Figure 1) [38], could also limit the experimenters' role while providing the important social presence and emotional feedback in a more reproducible form [39].
In conclusion, social presence and emotional feedback are meant to increase the effort, motivation and engagement of the participants throughout the learning. As any feedback, they must be carefully studied as they can be double-edged. On the one hand, they can benefit the learning outcome, depending on the participants' profile [6, 26, 33, 38]. On the other hand, as any feedback, they can have a detrimental impact on the user training and the reliability of experimental results when they are incorrectly designed and assessed [54].

Acknowledgements

We would like to thank all the participants and the experimenters, i.e., Aurélien Appriou, Camille Benaroch and Damien Caselli, for dedicating the time to conduct some of the experiments.

References

[1] Biasiucci, A., Leeb, R., Iturrate, I., Perdikis, S., Al-Khodairy, A., Corbet, T., Schnider, A., Schmidlin, T., Zhang, H., Bassolino, M., et al. (2018). Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. *Nature communications*, 9(1):1–13.

[2] Birbaumer, N. (2006). Breaking the silence: brain–computer interfaces (BCI) for communication and motor control. *Psychophysiology*, 43(6):517–532.

[3] Bismuth, J., Vialatte, F., and Lefaucheur, J.-P. (2020). Relieving peripheral neuropathic pain by increasing the power-ratio of low-β over high-β activities in the central cortical region with EEG-based neurofeedback: study protocol for a controlled pilot trial (SMRPain study). *Neurophysiologie Clinique*.

[4] Blankertz, B., Sannelli, C., Halder, S., Hammer, E. M., Kübler, A., Müller, K.-R., Curio, G., and Dickhaus, T. (2010). Neurophysiological predictor of SMR-based BCI performance. *Neuroimage*, 51(4):1303–1309.

[5] Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Muller, K. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal processing magazine*, 25(1):41–56.

[6] Bonnet, L., Lotte, F., and Lécuyer, A. (2013). Two brains, one game: design and evaluation of a multiuser BCI video game based on motor imagery. *IEEE Transactions on Computational Intelligence and AI in games*, 5(2):185–198.

[7] Cattell, R. and P. Cattell, H. (1995). Personality structure and the new fifth edition of the 16PF. *Educational and Psychological Measurement*, 55(6):926–937.

[8] Chapman, C., Benedict, C., and Schiöth, H. (2018). Experimenter gender and replicability in science. *Science advances*, 4(1):e1701427.

[9] Clerc, M., Bougrain, L., and Lotte, F. (2016). *Brain-Computer Interfaces 2: Technology and Applications*. John Wiley & Sons.

[10] Fatourechi, M., Bashashati, A., Ward, R. K., and Birch, G. E. (2007). EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical neurophysiology*, 118(3):480–494.

[11] Hammer, E. M., Halder, S., Blankertz, B., Sannelli, C., Dickhaus, T., Kleih, S., Müller, K.-R., and Kübler, A. (2012). Psychological predictors of SMR-BCI performance. *Biological psychology*, 89(1):80–86.

[12] Hattie, J. and Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1):81–112.

[13] Jaumard-Hakoun, A., Chikhi, S., Medani, T., Nair, A., Dreyfus, G., and Vialatte, F.-B. (2017). An apparatus to investigate western opera singing skill learning using performance and result biofeedback, and measuring its neural correlates. *Interspeech*.

[14] Jeunet, C., Jahanpour, E., and Lotte, F. (2016a). Why standard brain-computer interface (BCI) training protocols should be changed: an experimental study. *Journal of neural engineering*, 13(3):036024.

[15] Jeunet, C., N'Kaoua, B., and Lotte, F. (2016b). Advances in user-training for mental-imagery-based BCI control: Psychological and cognitive factors and their neural correlates. In *Progress in brain research*, volume 228, pages 3–35. Elsevier.

[16] Jeunet, C., N'Kaoua, B., Subramanian, S., Hachet, M., and Lotte, F. (2015). Predicting mental imagery-based BCI performance from personality, cognitive profile and neurophysiological patterns. *PloS one*, 10(12):e0143962.

[17] Jin, J., Miao, Y., Daly, I., Zuo, C., Hu, D., and Cichocki, A. (2019). Correlation-based channel selection and regularized feature optimization for mi-based bci. *Neural Networks*, 118:262–270.

[18] Kline, J., Blackhart, G., and Joiner, T. (2002). Sex, lie scales, and electrode caps: An interpersonal context for defensiveness and anterior electroencephalographic asymmetry. *Personality and Individual Differences*, 33(3):459–478.

[19] Kober, S. E., Witte, M., Ninaus, M., Neuper, C., and Wood, G. (2013). Learning to

modulate one's own brain activity: the effect of spontaneous mental strategies. *Frontiers in human neuroscience*, 7:695.

[20] Lécuyer, A. (2016). BCIs and Video Games: State of the Art with the OpenViBE2 Project. *Brain–Computer Interfaces 2: Technology and Applications*, pages 85–99.

[21] Levine, F. and De Simone, L. (1991). The effects of experimenter gender on pain report in male and female subjects. *Pain*, 44(1):69–72.

[22] Linn, M. and Petersen, A. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child development*, pages 1479–1498.

[23] Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., and Yger, F. (2018). A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005.

[24] Lotte, F. and Jeunet, C. (2018). Defining and quantifying users' mental imagery-based BCI skills: a first step. *Journal of neural engineering*, 15(4):046030.

[25] Lotte, F., Larrue, F., and Mühl, C. (2013). Flaws in current human training protocols for spontaneous brain-computer interfaces: lessons learned from instructional design. *Frontiers in human neuroscience*, 7.

[26] Mathiak, K. A., Alawi, E. M., Koush, Y., Dyck, M., Cordes, J., Gaber, T., Zepf, F., Palomero-Gallagher, N., Sarkheil, P., Bergert, S., et al. (2015). Social reward improves the voluntary control over localized brain activity in fMRI-based neurofeedback training. *Frontiers in behavioral neuroscience*, 9:136.

[27] McFarland, D. J. and Wolpaw, J. R. (2018). Brain–computer interface use is a skill that user and system acquire together. *PLoS biology*, 16(7):e2006719.

[28] Miyazaki, A. and Taylor, K. (2008). Researcher interaction biases and business ethics research: Respondent reactions to researcher characteristics. *Journal of Business Ethics*, 81(4):779–795.

[29] Morone, G., Pisotta, I., Pichiorri, F., Kleih, S., Paolucci, S., Molinari, M., Cincotti, F., Kübler, A., and Mattia, D. (2015). Proof of principle of a brain-computer interface

approach to support poststroke arm rehabilitation in hospitalized patients: design, acceptability, and usability. *Archives of physical medicine and rehabilitation*, 96(3):S71–S78.

[30] Neuper, C. and Pfurtscheller, G. (2010). *Brain-Computer Interfaces*, chapter Neurofeedback Training for BCI Control, pages 65–78. The Frontiers Collection.

[31] Neuper, C., Scherer, R., Reiner, M., and Pfurtscheller, G. (2005). Imagery of motor actions: Differential effects of kinesthetic and visual–motor mode of imagery in single-trial EEG. *Cognitive brain research*, 25(3):668–677.

[32] Nichols, A. and Maner, J. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of general psychology*, 135(2):151–166.

[33] Nijboer, F., Furdea, A., Gunst, I., Mellinger, J., McFarland, D., Birbaumer, N., and Kübler, A. (2008). An auditory brain–computer interface (BCI). *J Neur Meth*.

[34] Ono, T., Kimura, A., and Ushiba, J. (2013). Daily training with realistic visual feedback improves reproducibility of event-related desynchronisation following hand motor imagery. *Clinical Neurophysiology*, 124(9):1779–1786.

[35] Pfurtscheller, G. and Neuper, C. (2001). Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*, 89(7):1123–1134.

[36] Pichiorri, F., Morone, G., Petti, M., Toppi, J., Pisotta, I., Molinari, M., Paolucci, S., Inghilleri, M., Astolfi, L., Cincotti, F., et al. (2015). Brain–computer interface boosts motor imagery practice during stroke recovery. *Annals of neurology*, 77(5):851–865.

[37] Pillette, L. (2019). *Redefining and Adapting Feedback for Mental-Imagery based Brain-Computer Interface User Training to the Learners' Traits and States*. PhD thesis, Université de Bordeaux.

[38] Pillette, L., Jeunet, C., Mansencal, B., N'kambou, R., N'Kaoua, B., and Lotte, F. (2020). A physical learning companion for Mental-Imagery BCI User Training. *International Journal of Human-Computer Studies*, 136:102380.

[39] Pillette, L., Jeunet, C., N'Kambou, R., N'Kaoua, B., and Lotte, F. (2018). Towards artificial learning companions for mental imagery-based brain-computer interfaces. In *Workshop sur les "Affects, Compagnons Artificiels et Interactions"(ACAI)*, pages 1–8.

[40] Ramoser, H., Muller-Gerking, J., and Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446.

[41] Renard, Y., Lotte, F., Gibert, G., Congedo, M., Maby, E., Delannoy, V., Bertrand, O., and Lécuyer, A. (2010). OpenViBE: An open-source software platform to design, test and use brain-computer interfaces in real and virtual environments. *Presence: teleoperators and virtual environments*, 19(1):35–53.

[42] Roc, A., Pillette, L., Mladenovic, J., Benaroch, C., N'Kaoua, B., Jeunet, C., and Lotte, F. (2020). A review of user training methods in brain computer interfaces based on mental tasks. *Journal of Neural Engineering*.

[43] Roc, A., Pillette, L., N'Kaoua, B., and Lotte, F. (2019). Would motor-imagery based BCI user training benefit from more women experimenters? In *8th International BCI Conference*, pages 1–7.

[44] Rosenthal, R. (1963). On the social psychology of the psychological experiment: 1, 2 the experimenter's hypothesis as unintended determinant of experimental results. *American Scientist*, 51(2):268–283.

[45] Rosnow, R. and Rosenthal, R. (1997). *People studying people: Artifacts and ethics in behavioral research*. WH Freeman.

[46] Sexton, C. (2015). The overlooked potential for social factors to improve effectiveness of brain-computer interfaces. *Frontiers in systems neuroscience*, 9:70.

[47] Sollfrank, T., Ramsay, A., Perdikis, S., Williamson, J., Murray-Smith, R., Leeb, R., Millán, J., and Kübler, A. (2016). The effect of multimodal and enriched feedback on SMR-BCI performance. *Clinical Neurophysiology*, 127(1):490–498.

[48] Spencer, S. J., Steele, C. M., and Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, 35(1):4–28.

[49] Thomas, E., Dyson, M., and Clerc, M. (2013). An analysis of performance evaluation for motor-imagery based BCI. *Journal of neural engineering*, 10(3):031001.

[50] Tyson, P. D. (1982). The choice of feedback stimulus can determine the success of alpha feedback training. *Psychophysiology*, 19(2):218–230.

[51] Vandenberg, S. G. and Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and motor skills*, 47(2):599–604.

[52] Vourvopoulos, A. T., Jorge, C., Abreu, R., Figueiredo, P., Fernandes, J.-C., and Bermúdez i Badia, S. (2019). Efficacy and brain imaging correlates of an immersive motor imagery bci-driven vr system for upper limb motor rehabilitation: A clinical case report. *Frontiers in Human Neuroscience*, 13:244.

[53] Wheeler, S. C. and Petty, R. E. (2001). The effects of stereotype activation on behavior: a review of possible mechanisms. *Psychological bulletin*, 127(6):797.

[54] Wood, G. and Kober, S. (2018). EEG Neurofeedback Is Under Strong Control of Psychosocial Factors. *Applied psychophysiology and biofeedback*, 43(4):293–300.

Appendix

A Details regarding the analyses on the potential influence of artefact sources

Because brain signals are really small in amplitude and EEG suffers from very low signal to noise ratio (SNR), i.e., high vulnerability to artefact sources, we controlled for the most common artefact sources, i.e., electrooculography (EOG) and electromyography (EMG) [10]. The aim was to check if specific patterns could be found in EOG or EMG signals that could have affected MI classification by the BCI. The presence of such task-specific patterns could have confounded the measured MI-BCI performances. We thus wanted to assess how much EMG or EOG artefacts could have affected the recorded EEG signals and influenced the MI-BCI classification output and accuracy. To do so, we computed two metrics per source of potential artefacts.

First, we looked at left vs right MI classification accuracy, i.e., TAcc and EAcc, based on EOG or EMG signals, using a classifier built on the calibration runs. This was computed using CSP/LDA calibrated on the EOG or EMG signals only from the two calibration runs, filtered in the participant-specific discriminant frequency band. Note that we used the same frequency band as for the online experiment since only task-related EMG and EOG variations occurring in the same frequency band as the one used by the EEG-BCI classifier could have affected this classifier output, and therefore the resulting BCI accuracy. The resulting classifier was then applied on the subsequent runs to obtain a measure of EOG or EMG accuracy per run. The accuracies based on such calibration run can reflect the presence of task-specific EMG or EOG artefacts in EEG signals, during both the calibration and the training phases, which might have influenced online EEG-based BCI performances.

Second, the run-specific left vs right MI EOG or EMG accuracies were computed using a cross-validation method. EOG or EMG data only from each run, filtered in the participant-specific discriminant frequency band, were divided into five subsets of data. The CSP and an LDA were successively calibrated on four sets and tested on the remaining one. The run-specific EOG or EMG metric is the mean classification accuracy obtained for the five subsets for each of the runs. The run-specific accuracies reflect the presence of task-specific EOG or

EMG artefacts that could have affected online EEG-based BCI performance, during each run. The results of these analyses are presented in the sections below.

A.1 Checking the influence of EMG artefacts

We first assessed whether EMG artefacts, or real unsolicited hand movements from our participants, could have had an impact on our main results, i.e., the interactions we found between the evolution of trial-wise accuracy and experimenters' and participants' gender that we obtained with an EEG-based classification accuracy.

We inspected the potential relation between mean EEG-based classification accuracies, i.e., TAcc and EAcc, and EMG-based classification accuracies, i.e., calibration runs based and run specific, by performing analyses of correlation. We did not find any correlation between the mean calibration runs based EMG accuracy and the mean TAcc [Spearman correlation, r(54) = -.2, p = .15] nor with the mean EAcc [Spearman correlation, r(52) = -.15, p = .29]. No correlation could be found either between the mean run specific EMG accuracy and the TAcc [Spearman correlation, r(53) = -.1, p = .49] nor the EAcc [Spearman correlation, r(51) = -.86, p = .55].

A.2 Checking the influence of EOG artefacts

Similarly to the previous section, we inspected if EOG artefacts or eye movements performed by our participants could have had an impact on our main results that we obtained with EEG-based classification accuracies.

We inspected the potential relation between mean performances, i.e., TAcc and EAcc, and EOG-based classification accuracies, i.e., calibration runs based and run specific, by performing analyses of correlation. We did not find any correlation between the mean calibration runs based EOG accuracy and the mean TAcc [Spearman correlation, r(54) = -.23, p = .11] nor with the mean EAcc [Spearman correlation, r(52) = -.17, p = .22]. Though, a significant correlation could be found between the mean run specific EOG accuracy and both the TAcc [Spearman correlation, r(56) = .31, p = .02] and the EAcc [Spearman correlation, r(54) = .36, $p < 10^{-2}$].

We hypothesized that these significant correlations resulted from EEG acquisitions from the electrodes positioned to measure EOG. Indeed, when the same analysis was performed us-

ing cross-validation on data filtered on EOG frequency band, i.e., 0.5-4Hz, we did not find any correlation with the mean TAcc [Spearman correlation, r(54) = .05, p = .73] nor with the mean EAcc [Spearman correlation, r(52) = .12, p = .39].

B Details regarding the analyses on the potential influence of MRS and autonomy differences in participant groups

As stated in Section 3 -Results-, the groups of participants formed using the participants' and experimenters' gender had differences in terms of mental rotation scores and autonomy. Therefore, we studied the potential impact of these differences on the results presented in Section 3.1 -H1 - MI-BCI performances-.

We ran our same main analyses than in this section (two 3-way repeated measures mixed ANOVAs with *"ExpGender"*, *"ParGender"* and *"Run"* as independent variables and the repeated measures of performance over the runs, i.e., TAcc or EAcc, as dependent variable) using the autonomy, i.e., *"Autonomy"*, or the mental rotation score, i.e., *"MRS"*, of the participants as covariate. When performing the analysis on the TAcc we found no impact of the autonomy (*"Autonomy"* [F(1, 51) = 0.26, p = .61, $\eta^2 < 10^{-2}$], *"Autonomy*Run"* [F(2.48, 126.6) = 0.81, p = .47, η^2 = .02]) nor of the mental rotation score (*"MRS"* [F(1, 51) = 1.75, p = .19, η^2 = .03], *"MRS*Run"* [F(2.47, 125.79) = 1.52, p = .22, η^2 = .03]). When investigating the EAcc we did not find any single effect or interaction of the autonomy (*"Autonomy"* [F(1, 51) = 0.44, p = .51, $\eta^2 = 10^{-2}$], *"Autonomy*Run"* [F(2.1, 107.14) = 1.46, p = .24, η^2 = .03]) nor of the mental rotation score (*"MRS"* [F(1, 51) = 1.05, p = .31, η^2 = .02], *"MRs*Run"* [F(2.18, 111.18) = 1.35, p = .27, η^2 = .03]) either.

C Details regarding the analyses on the potential influence of experimenters' gender on the user-experience

We analysed the influence of experimenters' and participants' gender on the five dimensions of the user-experience, i.e., mood, mindfulness, motivation, cognitive load and sense of agency. First, we checked if the performances had an impact on the reported user-experience measures. We found that both the TAcc [Spearman correlation, r(56) = .38, $p < 10^{-2}$] and EAcc

[Spearman correlation, r(56) = .34, p = .01] metrics were positively correlated to the sense of agency post training.

Therefore, we performed five 2-way ANOVAs or ANCOVAs, one per dimension, with *"ExpGender*ParGender"* as independent variables and either the measure of cognitive load, sense of agency, mood, mindfulness or motivation as dependent variable. Performances averaged over all runs, i.e., TAcc or EAcc, were used as covariate if they were correlated to the dependent variable.

No influence was found on the cognitive load reported post training of *"ExpGender"* [F(1, 52) = 1.65, p = .2, η^2 = .03], *"ParGender"* [F(1, 52) = 2.89, p = .1, η^2 = .05] nor *"ExpGender*ParGender"* [F(1, 52) = 0.05, p=0.95, $\eta^2 < 10^{-3}$].

No influence was found on the sense of agency of *"ExpGender"* [F(1, 52) = 0.03, p = .85, $\eta^2 = 10^{-3}$], *"ParGender"* [F(1, 52) = 0.01, p = .92, $\eta^2 < 10^{-3}$] nor *"ExpGender* ParGender"* [F(1, 56) = 0.44, p = .51, $\eta^2 < 10^{-2}$] using the TAcc as covariable. Neither was there any influence found with the EAcc as covariable of *"ExpGender"* [F(1, 56) = 0.08, p = .78, $\eta^2 = 10^{-2}$], *"ParGender"* [F(1, 52) = 10^{-3}, p = .97, $\eta^2 < 10^{-3}$] nor *"ExpGender* ParGender"* [F(1, 52) = 0.52, p = .47, η^2 = .01].

No influence was found on the difference of mood reported post and pre training of *"ExpGender"* [F(1, 52) = 0.06, p = .81, $\eta^2 = 10^{-3}$], *"ParGender"* [F(1, 52) < 10^{-2}, p = .93, $\eta^2 < 10^{-3}$] nor *"ExpGender*ParGender"*[F(1, 52) = 0.13, p = .72, $\eta^2 < 10^{-2}$].

No influence was found on the difference of mindfulness reported post and pre training of *"ExpGender"* [F(1, 52) = 0.04, p = .85, $\eta^2 = 10^{-3}$] or *"ExpGender*ParGender"* [F(1, 52) = 0.92, p = .34, η^2 = .02]. Though, a significant impact of *"ParGender"* [F(1, 52) = 6.23, p = .02, η^2 = .11] was found. Overall, men participants had a decrease of mindfulness (*MmindfulnessMen* = -8.33, *SD* = 3.01) whereas women participants had an increase (*MmindfulnessWomen* = 2.5, *SD* = 3.12) of mindfulness over the session.

No influence was found on the difference of motivation reported post and pre training of *"ExpGender"* [F(1, 52) = 0.63, p = .43, η^2 = .01], *"ParGender"* [F(1, 52) = 0.78, p = .38, η^2 = .02] nor *"ExpGender*ParGender"* [F(1, 52) = 0.97, p = .33, η^2 = .02].